Logistic Regression

Types of Regression

Regression analysis is a predictive modelling technique. It estimates the relationship between a dependent (target) and an independent variable(predictor)

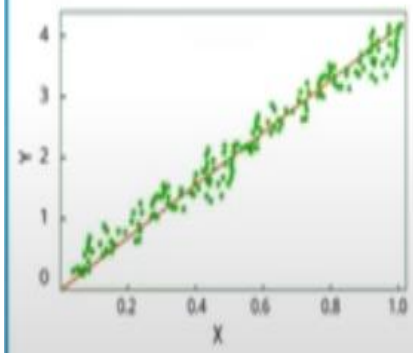
There are three types of regression

1. Linear Regression
2. Logistic regression
3. Polynomial regression

Types

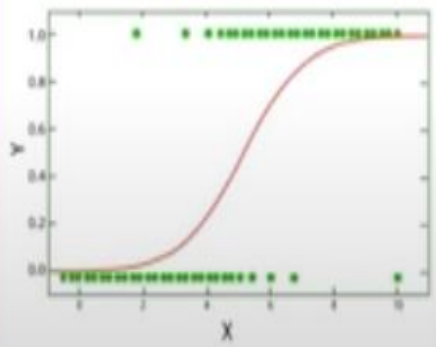
Linear Regression

- When there is a linear relationship between independent and dependent variables.



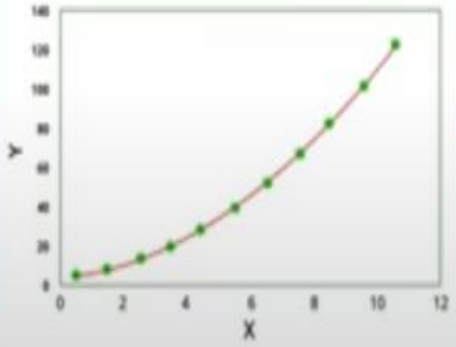
Logistic Regression

- When the dependent variable is categorical (0/ 1, True/ False, Yes/ No, A/B/C) in nature.



Polynomial Regression

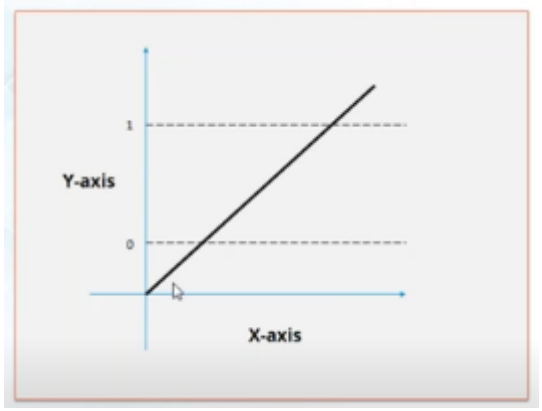
- When the power of independent variable is more than 1.

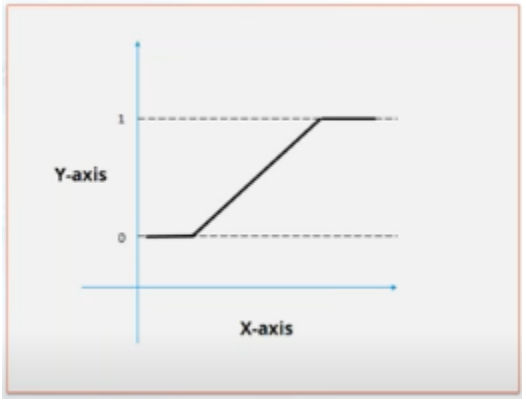


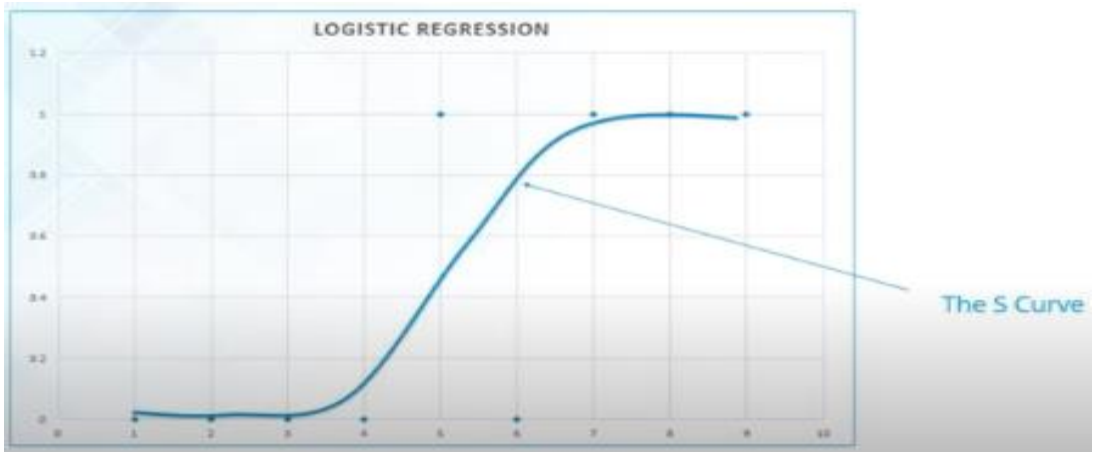
Need of logical Regression

1. The name logistic regression is used when the dependent variable is discrete. For example it has only two values, such as 0 and 1 or Yes and No or one of the choices A,B,C in case of multiple logistic regression.
2. logistic regression is also known as logit regression or logit model
3. As value of y should lie between 0 and 1, a linear line has to be clipped and S-curve is used.

S-curve for Logistic Regression







Formula for logistic regression

- If the variable y depends on independent var x, Equation of straight line is
- $y = b_0 + b_1 x$(range of y is -∞ to ∞)
- Consider the equation $y = b_0 + b_1 x$ for range of y from 0 to 1
- In such case, range of y/(1-y) is
- So range of log{y/(1-y) } is

Formula for logistic regression

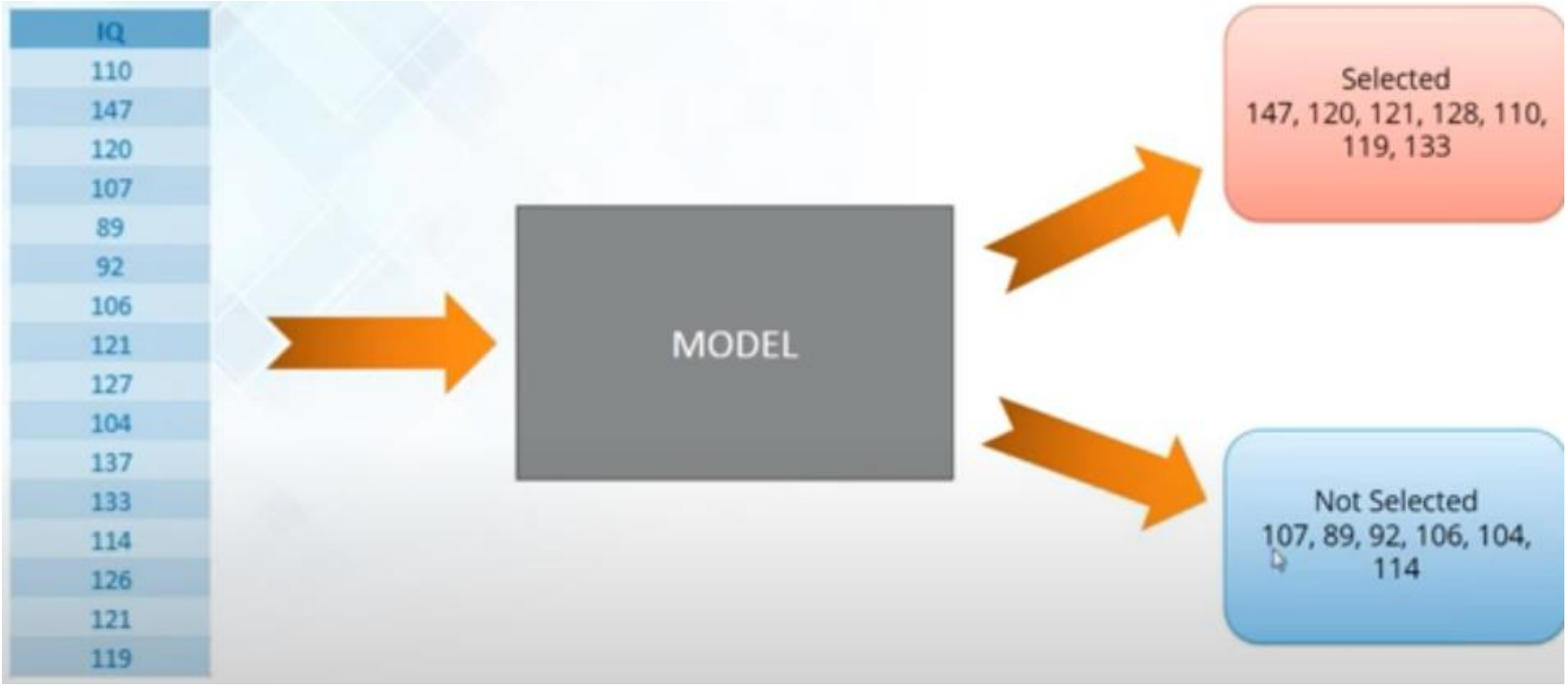
- If the variable y depends on independent var x, Equation of straight line is
- $y = b_0 + b_1 x$(range of y is -∞ to ∞)
- Consider the equation $y = b_0 + b_1 x$ for range of y from 0 to 1
- In such case, range of y/(1-y) is 0 to ∞
- So range of log{y/(1-y) } is -∞ to ∞
- Since y gives the probability, use the notation p and find the expression for p
- $p(x) = \frac{e^{b_0+b_1 x}}{1+e^{b_0+b_1 x}} = \frac{1}{1+e^{-(b_0+b_1 x)}}$ (Logistic Function)
- We should predict y=1 when p>= 0.5 and y=0 when p<0.5. This means guessing 1 whenever $b_0 + b_1 x$ is nonnegative and 0 otherwise.
- The decision boundary separating the two predicted classes is the solution of $b_0 + b_1 x = 0$
- $log\left(\frac{p}{1-p}\right) = b_0 + b_1 x$ **log - odds** or **odds ratio** or **logit** function

Table for p and logit(p) values

The following table shows the logit for various values of p.

P	Logit(P)	P	Logit(P)
0.001	-6.907	0.999	6.907
0.01	-4.595	0.99	4.595
0.05	-2.944	0.95	2.944
0.10	-2.197	0.90	2.197
0.20	-1.386	0.80	1.386
0.30	-0.847	0.70	0.847
0.40	-0.405	0.60	0.405
0.50	0.000		

Example to recruit people



Example

- For example, if we are modeling people's sex as male or female from their height, then the Y=1 could be male and the logistic regression model could be written as the probability of male given a person's height, or more formally:
- P(sex=male|height)
- Written another way, we are modeling the probability that an input (X) belongs to the default class (Y=1), we can write this formally as:
- P(X) = P(Y=1|X)

Ex: In logistic model people's sex as male or female from their height, then the Y=1 could be male , check whether a person with height 150 cm is a male . (given b0=-100, b1=0.6)

- $p(x) = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}} = \frac{1}{1+e^{-(b_0+b_1x)}}$

$$= 0.99995460213 > 0.5$$

Y=1

Hence a person with height 150 cm is a male .

Also it means that 99% of the time we will generate for the classification 1 and remaining 1% for classification 0

More about Logistic Regression

- In case of binary response, mean response is probability.
- Logistic Function is given by $p = \frac{1}{1+e^{-(b_0+b_1 x)}}$.
- The portion $b_0 + b_1 x$ is called linear predictor.
- The Odds ratio is designed to determine how the odds of success $\frac{p}{1-p}$ increases as certain changes in regressor value occur.

- Logit is log of odds and odds are function of p, the probability of 1.
- Log(odds)= logit (p) = $log\left(\frac{p}{1-p}\right)$
- The log Odds Ratio Transformation:

The difference between two log odds is used to compare two proportions such as male versus females.

$$l_1 - l_2 = \frac{log\left(\frac{p_1}{1-p_1}\right)}{log\left(\frac{p_2}{1-p_2}\right)}$$

Example: Comparing the proportions of female and male Instagram users

From the database for Instagram Users following data was analysised:

		User		
	Count Row %	No	Yes	Total
Sex	1Women	209 38.92	328 61.08	537
	2Men	298 56.02	234 43.98	532
	Total	507	562	1069

To use this in logistic regression, we need to use a numeric code. For our problem, we will use an indicator of whether or not the adult is a woman

$$x = \begin{cases} 1 & if\ the\ person\ is\ a\ woman \\ 0 & if\ a\ person\ is\ a\ man \end{cases}$$

For women, p

odds

Similarly, for men

Example: Comparing the proportions of female and male Instagram users

From the database for Instagram Users following data was analysised:

		User		
	Count Row %	No	Yes	Total
Sex	1Women	209 38.92	328 61.08	537
	2Men	298 56.02	234 43.98	532
	Total	507	562	1069

To use this in logistic regression, we need to use a numeric code. For our problem, we will use an indicator of whether or not the adult is a woman

$$x = \begin{cases} 1 & if\ the\ person\ is\ a\ woman \\ 0 & if\ a\ person\ is\ a\ man \end{cases}$$

For women, $p = \frac{378}{537} = 61.08\% = 0.6108$

odds

Similarly, for men

Example: Comparing the proportions of female and male Instagram users

From the database for Instagram Users following data was analysised:

		User		
	Count Row %	No	Yes	Total
Sex	1Women	209 38.92	328 61.08	537
	2Men	298 56.02	234 43.98	532
	Total	507	562	1069

To use this in logistic regression, we need to use a numeric code. For our problem, we will use an indicator of whether or not the adult is a woman

$$x = \begin{cases} 1 & if\ the\ person\ is\ a\ woman \\ 0 & if\ a\ person\ is\ a\ man \end{cases}$$

For women, $p = \frac{378}{537} = 61.08\% = 0.6108$

$$\text{odds} = \frac{p}{1-p} = \frac{0.6108}{1-0.6108} = 1.5694$$

Similarly, for men

Example: Comparing the proportions of female and male Instagram users

From the database for Instagram Users following data was analysised:

		User		
	Count Row %	No	Yes	Total
Sex	1Women	209 38.92	328 61.08	537
	2Men	298 56.02	234 43.98	532
	Total	507	562	1069

To use this in logistic regression, we need to use a numeric code. For our problem, we will use an indicator of whether or not the adult is a woman

$$x = \begin{cases} 1 & if\ the\ person\ is\ a\ woman \\ 0 & if\ a\ person\ is\ a\ man \end{cases}$$

For women, $p = \frac{378}{537} = 61.08\% = 0.6108$

$$\text{odds} = \frac{p}{1-p} = \frac{0.6108}{1-0.6108} = 1.5694$$

Similarly, for men we have $\text{odds} = \frac{p}{1-p} = \frac{0.4398}{1-0.4398} = 0.7851$

As we use linear regression, we use y for the response variable. So for women,

$$y = \log(odds) = \log(1.5694) = 0.4507$$

And for men,

$$y = \log(odds) = \log(0.7851) = -0.2419$$

In these expressions for the log odds, we use y as the observed value of the response variable, the log odds of using Instagram. We are now ready to build the logistic regression model. We model the log odds as a linear function of the explanatory variable:

$$log\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

For logistic regression, we use natural logarithms. There are tables of natural logarithms, and many calculators have a built-in function for this transformation.

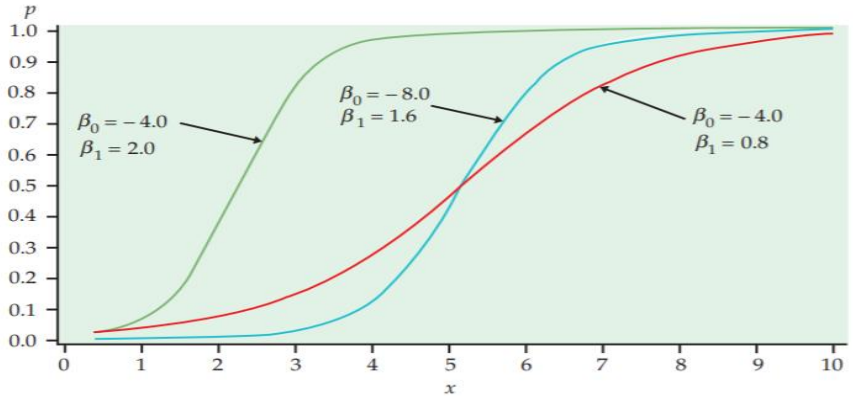


FIGURE 14.2 Plot of p versus x for different logistic regression models.

The statistical model for logistic regression is $log\left(\frac{p}{1-p}\right) = b_0 + b_1 x$

Where p is the binomial proportion and x is explanatory variable. The parameters of the logistic model are $b_0\ and\ b_1$.

For our Instragram example, there are n = 1069 young persons in the sample. The explanatory variable is sex, which we have coded using an indicator variable with values x = 1 for women and x = 0 for men. The response variable, y, is also an indicator variable. Thus, each person either is an Instagram user or is not an Instagram user. Think of a process of selecting a young person at random and recording y and x. The model says that the probability, p, that this person is an Instagram user can depend upon the user's sex (x = 1 or x = 0). So there are two possible values for p- $p_{women}\ and\ p_{men}$

For women, $log\left(\frac{p_{women}}{1-p_{women}}\right) = b_0 + b_1$ (as x=1 for women) =0.4507

For men, $log\left(\frac{p_{men}}{1-p_{men}}\right) = b_0$ (as x=0 for men) = -0.2419

Solve these two equations to get, $b_0 = -0.2419\ and\ b_1 = 0.6926$

The fitted logistic modal is

$\log(odds) = -0.2419 + 0.6926x$

The slope in this logistic regression model is the difference between the log odds for men and the log odds for women

Also we have $\frac{odds_{women}}{odds_{men}} = e^{0.6926} = 1.999$

In this case, we would say that the odds for women are about twice the odds for men. Notice that we have chosen the coding for the indicator variable so that the regression slope is positive. This will give an odds ratio that is greater than 1.

Had we coded men as 1 and women as 0, the sign of the slope would be reversed and the odds ratio would be 0.500. The odds for men are about half of the odds for women.

Statistical inference for logistic regression

- Statistical inference for logistic regression is very similar to statistical inference for simple linear regression. We calculate estimates of the model parameters and standard errors for these estimates
- The ratio of the estimate of the slope to the standard error is the basis for hypothesis tests
- A level C confidence interval for the slope b_1 is $b_1 \mp z * SE$
- A level C confidence interval for the odds ratio is $\left(e^{b_1 - z*SE}, e^{b_1 + z*SE}\right)$

Ex The output for the instragram example, has values 1 for women and 0 for men, the parameter estimates are $b_0 = -0.2419\ and\ b_1 = 0.6926$ The standard errors are 0.0873 and 0,1240 resp. Find 95% confidence interval for the slope and odds ratio.

- Solution
- Slope is b_1
- Value of z for 95% confidence interval is 1.96
- 95% confidence interval for the slope is $b_1 \mp z * SE$ = (0.44896, 0.93504)

- 95% confidence interval for the odds ratio is $\left(e^{b_1 - z*SE}, e^{b_1 + z*SE}\right)$ = (1.57, 2.55)

exercise1

- the relative risk of developing cardiovascular disease (CVD) for people with low- and high-salt diets was estimated.

	Salt in diet		
Developed CVD	Low	High	Total
Yes	88	112	200
No	1081	1134	2215
Total	1169	1246	2415

(a) For each salt level, find the probability of developing CVD.

(b) Convert each of the probabilities that you found in part (a) to odds.

(c) Find the log of each of the odds that you found in part (b).

exercise2

- A survey was conducted for some students to find the number of hours each student spent in daily studying and whether they passed or failed. Using Logistic regression, it was found that

  Log(odds of passing exam)= 1.5046.hours-4.0777

  Find

(i) Odds of passing the exam

(ii) Probability of passing exam

(iii) Slope of odds ratio

(iv) Probability of passing exam if a student studies 2 hrs daily.

Loss Function

- A Loss Function is a measure of fit between a Mathematical model of data and the actual data.
- We choose the parameters of the model to minimise the badness of fit or to maximise the goodness of fit of the model of the data.
- Residual Analysis detects outliers. Identifies influential observations and diagnoses the appropriateness of the logistic model.

Training set

Size of tumor	Malignant?
0.1	0
0.3	0
0.7	0
0.9	1
1.4	1
1.7	1
2.3	1

- This is a **binary classification** problem as each tumor is either Malignant (y = 1) or Benign (y = 0). These are the 2 classes here. Our aim is to **come up with a probability function** that takes in an input X (size of tumor) and return '***what is the probability of this tumor to be malignant***'.

- Since, **probability of any event to happen is [0,1] (between 0 and 1, including both), this function definitely seems fit to be used as a probability function for logistic regression.**

- **what is Cost/error function**.
- Suppose you were able to come up with a probability function. You fed it X(size of tumor) = 0.9 and it gave probability for it to be malignant = 0.3, which means it has more chance of being benign. But clearly from our training set this definitely is wrong as for X = 0.9, Y = 1 i.e malignant. So t**his is an error.**

•Logistic Regression is used to predict categorical variables. Binary(yes/no) where the variable has one of 2 possible categories, or Multinomial where there can be more than 2 categories.
•Say you have patient data - age, gender, result of blood tests, size of some tumor. You want to predict whether the tumor is cancerous or not.
•Or you have the scores of students in various subjects and tests, along with their age and other statistics and you want to see if the student would be admitted to a particular university or not. This is an application of binary regression.

•In the latter example, if you try to regress and find the major the student is most likely to take, that would be multinomial regression.

- Life insurance actuaries use logistic regression to predict, based on given data on a policy holder (e.g. age, gender, results from a physical examination) the chances that the policy holder will die before the term of the policy expires.
- Political campaigns try to predict the chances that a voter will vote for their candidate (or do something else desirable, such as donate to the campaign).
- Bankers use it to predict the chances that a loan applicant will default on the loan.
- Marketers use it to predict whether a customer will respond to a particular ad (whether by clicking on a link or sending back a self-enclosed mailer).
- Weather forecasters use it to predict the "chance of rain" you see every morning.